Cox regression survival analysis with compositional covariates: application to modelling mortality risk from 24-h physical activity patterns

McGregor, D.E.; Palarea-Albaladejo, J.; Dall, P.M.; Hron, K.; Chastin, S.F.M.

Published in:
Statistical Methods in Medical Research

DOI:
10.1177/0962280219864125

Publication date:
2020

Document Version
Author accepted manuscript

Title: Cox regression survival analysis with compositional covariates: application to modelling mortality risk from 24-hour physical activity patterns.

Authors:

McGregor DE[1,2], Palarea-Albaladejo J[2], Dall PM[1], Hron K[3], Chastin SFM[1,4]

Affiliations:

1 School of Health and Life Science, Glasgow Caledonian University, Glasgow, Scotland, UK

2 Biomathematics and Statistics Scotland, Edinburgh, Scotland, UK

3 Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc, Olomouc, Czech Republic

4 Department of Movement and Sport Science, Ghent University, Ghent, Belgium

Corresponding author:

Duncan E McGregor, Glasgow Caledonian University, 70 Cowcaddens Road, Glasgow, G4 0BA, Scotland, UK

Abstract

Survival analysis is commonly conducted in medical and public health research to assess the association of an exposure or intervention with a hard end outcome such as mortality. The Cox (proportional hazards) regression model is probably the most popular statistical tool used in this context. However, when the exposure includes compositional covariables (that is, variables representing a relative makeup such as a nutritional or physical activity behaviour composition), some basic assumptions of the Cox regression model and associated significance tests are violated. Compositional variables involve an intrinsic interplay between one another which precludes results and conclusions based on considering them in isolation as is ordinarily done. In this work, we introduce a formulation of the Cox regression model in terms of log-ratio coordinates which suitably deals with the constraints of compositional covariates, facilitates the use of common statistical inference methods, and allows for scientifically meaningful interpretations. We illustrate its practical application to a public health problem: the estimation of the mortality hazard associated with the composition of daily activity behaviour (physical activity, sitting time and sleep) using data from the U.S. National Health and Nutrition Examination Survey (NHANES).

Keywords: survival analysis, Cox regression, compositional data, time use, accelerometry, physical activity, sedentary behaviour, NHANES

1. Introduction

Statistical survival analysis methods are commonly used in medical and public health studies where the outcome of interest is time to a specific event (1). This event is often a hard end outcome such as death, relapse of a disease or development of a new disease. For example, in public health, scientists might be interested in quantifying the risk of mortality to being exposed to specific behavioural or environmental factors over time (2). Similarly, clinical trials are conducted to assess the efficacy of interventions or new treatment regimes and the risk of potential adverse effects. These studies involve following participants for a long time and recording time to the event of interest. Cox's proportional hazards regression analysis (3) is one of the most common statistical modelling tools used to analyse such types of data.

The Cox regression model simplifies the analysis of survival rates by defining an instantaneous rate of mortality (or some other event) referred to as the hazard function, and estimating the proportional difference in the hazard function either between treatment groups or associated with changes in the exposure variables (4).

The hazard function h is formally defined with reference to the probability of survival as

$$S(t) = \exp\left(-\int_0^t h(u)du\right) \qquad [1]$$

where $S(t)$ is the probability of survival at time t. The standard Cox model specifies that the hazard function at time t is

$$h(t;\boldsymbol{w}) = h_0(t)\exp(\boldsymbol{\alpha}^T\boldsymbol{w}),$$

where $h_0(t)$ is an unspecified baseline hazard function, $\boldsymbol{w}$ is the vector of explanatory variables (which are zero-valued for some predefined reference strata), and $\boldsymbol{\alpha}$ is the corresponding vector of coefficients to be fitted to the data[1]. It is then simple to compare any two individuals 1 and 2 in terms of their expected hazards, given their respective set of covariate values, by using the hazard ratio $h_1(t)/h_2(t)$, which avoids defining $h_0(t)$ explicitly as it cancels out in the ratio and does not

[1] Note that we consider column vector throughout of this work and the superscript T refers to the matrix transpose operation.

depend on time. Thus, the hazard is proportional over time. The Cox regression model is usually expressed in terms of the logarithm of the hazard relative to the baseline as

$$y = \ln\left(\frac{h(t;\boldsymbol{w})}{h_0(t)}\right) = \boldsymbol{\alpha}^T\boldsymbol{w} \qquad [2]$$

which turns it into a linear function on the predictors. Basic assumptions of the model are that the associations between the covariates and hazard rate are not time-dependent, that there is no multicollinearity among covariates, and that their sample space is the real space endowed with Euclidean geometry. However, a key issue when the exposure variables are compositional is that they are intrinsically co-dependent variables carrying relative information and naturally defined on a simplex as representation of their sample space. These are features not accounted for by the ordinary Cox regression model.

Compositional data are common in many disciplines, including medical and public health research. Some recent discussions and applications of compositional methods in the area include nutritional epidemiology (5), health care research (6), microbiome and next-generation sequencing studies (7,8), and physical activity epidemiology (9,10). The key fact is that the measured values of a part of a composition are only meaningful when they are put in contrast to the values of the other parts or components. Importantly, the results and conclusions should be the same regardless of the chosen scale (what is referred to as the scale invariance property). Moreover, when the data are closed (i.e. represented with an arbitrary but fixed sum of the parts, commonly 1 or 100), a multicollinearity problem arises in regression analysis as a consequence of the singularity of the raw covariance matrix between compositional variables, which was already recognised in early approaches to mixture experiments (11–14). Although for model fitting purposes this could be technically overcome using e.g. Moore-Penrose pseudo-inverse matrix, this does not guarantee the quality of the regression estimation, and does not allow the investigation of the relative importance and role of the parts of the composition from the model coefficients (15). If the special nature of compositions is not considered adequately, the relationships between parts and inferences about any of them will be dependent on the presence or absence of other parts, what is known as the

subcompositional incoherence problem (16). This is, for example, highly relevant in physical activity research, where analyses of the allocation of time across daily activities considering either the entire 24-hour day or the waking day only (excluding sleep time) might lead to conflicting conclusions when both data sets are represented, e.g. in percentages. The log-ratio methodology for compositional data analysis (16) provides a coherent statistical framework which resolves these issues through the use of log-ratio type representations of the compositional variables. Although approaches have been presented for ordinary regression modelling with compositions (9,16–20), and survival analysis has been applied in respect of compositional outcomes such as the composition of different strains of bacteria (12), to our knowledge survival analysis based on the Cox regression model, or indeed survival analysis in general, with compositional explanatory variables remains formally unexplored.

In the following sections we introduce and discuss Cox regression analysis and related inference for the case of compositional exposure variables. We then illustrate the methodology by applying it to model the relationship between mortality and the composition of time spent in sleep, physical activity and sedentary behaviour using U.S. National Health And Nutrition Examination Survey (NHANES) data.

2. The compositional Cox regression model

The formal definition of compositional data has evolved from the particular case of multivariate positive data subject to a fixed sum constraint, as for example the distribution of the use of time over the 24 hour day cycle (either expressed in percentages adding up to 100 or in hours per day), to a more general characterisation as interdependent data with parts carrying relative information with respect to each other, with the observations not necessarily adding up to a same constant value (16). Thus, a composition $\boldsymbol{x} = (x_1, \dots, x_D)^T$ consisting of D parts represents an equivalence class in a $(D-1)$-dimensional sample space (21), where the data can be expressed in different (equivalent) relative scales (say percentages, proportions, hours/day, and so on) by applying a multiplicative

factor while the relative information remains the same. The common choice of normalising (or closing) the data and expressing them in percentages adding up to 100 is then just one of the possible equivalent representations.

2.1 Defining log-ratio coordinates

In order to deal with the particularities of compositional data in our survival analysis context, we adhere to the well-established methodological approach originated from the seminal work by Aitchison (22) and based on working with real-valued log-ratios between parts of the composition. Statistical results obtained using log-ratio coordinates can be then transferred back to the simplex to be represented in terms of the original composition. There have been a number of proposals to constructing such log-ratios and we briefly review them in the following subsections.

2.1.1 Additive log-ratios

For modelling purposes, Aitchison (22) originally proposed the so-called additive log-ratio (alr) transform, so that a set of $D-1$ transformed covariates y_i, consisting of the log-ratios of parts x_i with respect to one part x_c,

$$y_i = ln\left(\frac{x_i}{x_c}\right) \text{ for } i \neq c,$$

serve as a substitute for the original composition in fitting a statistical model by ordinary procedures. A recent review of compositional data analysis advocating for this approach can be found in (23).

2.1.2 Isometric log-ratios

Recent advances in the geometric characterization of the simplex as representation of the sample space of compositional data, namely a simplex S^D consisting of D-part compositions with its own Euclidean space structure (24), allow the definition of isometric log-ratio (ilr) coordinates which express the original composition $\boldsymbol{x} \in S^D$ in real coordinates in the ordinary real space $\mathbb{R}^{D-1}$.

Notably, due to orthonormality of the ilr-coordinates, distances or other measures of differences between compositions are preserved by their log-ratio counterparts. This is not the case with the alr mapping though, which projects compositions onto an oblique real coordinate system.

A procedure known as sequential binary partition (SBP) can be used to construct tailored ilr-coordinates, usually called compositional balances (z_i, for $i = 1, \dots, D-1$), representing log-contrasts, i.e. linear combinations $a_1 \ln x_1 + \dots + a_D \ln x_D$, with $a_1 + \cdots + a_D = 0$, between subsets of parts of the composition $\boldsymbol{x}$ (25). This permits ilr-coordinates to be defined with reference to domain knowledge or to an initial investigation of the co-variation structure of the data (e.g. grouping parts which are highly proportional). Compositional balances are obtained using SBP by successive splits of the parts of the composition $\boldsymbol{x}$ into 2 mutually exclusive groups until only groups of 1 part are left. These two groups are denoted below by the indices $+$ and $-$. The collection of balances z_i, for $i = 1, \dots, D-1$, is obtained as

$$z_i = \sqrt{\frac{r_i s_i}{r_i + s_i}} \ln \left[\frac{\left(\prod_{k=1}^{r_i} x_{i_k}^{+} \right)^{1/r_i}}{\left(\prod_{l=1}^{s_i} x_{i_l}^{-} \right)^{1/s_i}} \right], \qquad [3]$$

where $x_{i_k}^{+}$ and $x_{i_l}^{-}$ refer to the subsets of r_i and s_i components going, respectively, into the $+$ (numerator) and $-$ (denominator) groups. The $D-1$ balances fully represent the information in the composition $\boldsymbol{x}$, and are appropriate for use in standard statistical modeling. Note that the log-ratio term in Eq. [3] is computed as the ratio between the geometric means of the corresponding $+$ and $-$ components. The normalizing constant makes individual balances comparable and ensures orthonormality of the resulting coordinate system. The value of the balance can then be understood as a measure of the relative allocation between the two groups of parts of the composition, given in terms of the normalized difference between their logged geometric means (i.e. equivalently the normalized difference between the arithmetic means of the log-transformed parts). It is worth noting that regression models using either alr- or ilr-coordinates will be equivalent in that the predicted responses will be identical.

Note that there are infinite possible ilr-coordinate representations of a composition. However, these are all orthogonal rotations of each other and the fitted models from the same data will be entirely equivalent, with the only difference being obviously the value of the coefficients associated to the particular set of ilr-coordinates. The rotations can be defined with reference to a different orthogonal mapping $\boldsymbol{M}$ onto so-called centered log-ratio (clr) coefficients. The details of this process are described in previous work (16). For our purposes, it is sufficient to note that the clr coefficients, originally proposed by Aitchison (22), are uniquely defined as the set of log-ratios of the individual parts to the geometric mean of all them

$$\big(clr(\boldsymbol{x})\big)_i = \ln\left(\frac{x_i}{(\prod_{i=1}^{D} x_i)^{1/D}}\right),\ i = 1, \dots, D.$$

A direct consequence of the ability to define any ilr-coordinates with reference to an orthogonal mapping on uniquely defined clr-transformed parts, is that we can switch between two different partitions by an orthogonal mapping. For example, switching from a basis defined by $\boldsymbol{M}_1$ to a basis defined by $\boldsymbol{M}_2$ can be achieved through the rotation matrix

$$\boldsymbol{H} = \boldsymbol{M}_2\boldsymbol{M}_1^T . \qquad [4]$$

2.1.3 Pivot isometric log-ratios

The use of ilr-coordinates guarantees desirable formal properties like scale invariance, subcompositional coherence, and orthonormality (24). Given a composition which is relevant for the scientific question at hand, a particular system of ilr-coordinates obtained by SBP can be helpfully used to highlight in the first ilr-coordinate the relative importance or dominance of one part against the geometric average of the others in that particular composition of interest (26). Thus, without loss of generality, if we wished to investigate the relative importance of the part x_j against all other parts (e.g. one physical activity behavior versus all the others) we would express a composition $\boldsymbol{x} = (x_1, x_2, \dots, x_D)^T$ as a set of $D - 1$ ilr-coordinates including

$$z_1^j = \sqrt{\frac{D-1}{D}} \ln\left(\frac{x_j}{\left(\prod_{i \neq j} x_i\right)^{1/D-1}}\right) \quad [5]$$

as first ilr-coordinate. The remaining ilr-coordinates of the same set would then be defined with reference to the subcomposition excluding x_j. Note that the relative importance of any other part (with respect to the geometric mean of the others in the composition) can then be highlighted in the first ilr-coordinate z_1 by simply permuting the order of the parts in $\boldsymbol{x}$ to put the one of interest in first place, then generating an alternative but equivalent set of ilr-coordinates for the same composition (resulting from orthogonal rotation from the first one). This strategy has come to be called the pivot coordinate representation (26) and will be used later in Section 4. It is important to note that subcompositional coherence only guarantees that the same specific log-ratios will remain the same between a composition and its subcompositions. It does not imply that the first pivot coordinate of a composition and any subcomposition from it, or between any two subcompositions from it, will be the same. This is expectable as the geometric mean of the remaining parts in the denominator of the coordinate will obviously change with the change in the composition used as reference.

Pivot ilr-coordinates are particularly useful in first approaches to a problem with no a priori knowledge. They also allow the provision of results in a format which resembles common outputs from regression model fits, which eases the transition from ordinary to compositional modelling. The extra computational burden from actually fitting several regression models, one for each orthogonal rotation to assess the statistical significance of the corresponding first pivot coordinate, is actually not noticeable in practical settings using statistical software which automates such calculations.

Alternatively, the practitioner may prefer to define tailored ilr-coordinates generally through SBP as described in 2.1.2 for a more specific analysis, focusing on (and e.g. test statistical significance of) particular log-ratios of interest based on domain-specific knowledge or on a data-driven procedure (see e.g. (27, 36)). For example, if the parts of the composition can be meaningfully partitioned into two or more subsets, then it may be sensible to focus on balances between them. In the physical

activity context, one might wish to consider moderate and vigorous physical activity (MPA and VPA respectively) together by amalgamating them in relation to the subcomposition of all other activity types. In practice, VPA is absent from a large number of people's daily routines to an extent that makes imputation of time to this category unreliable, and the pragmatic decision in the area is often to combine time in any of these behaviors into MVPA. Notably, amalgamation of compositional parts is fully relevant within the log-ratio framework when it is done beforehand (23, 25).

Compositional data analysis based on ilr-coordinates has become in recent years the most popular choice in varied scientific areas. As a particular case, the characteristics of pivot ilr-coordinates have been important in our view to successfully introduce and popularize compositional analysis in physical activity research (10,20,28,29). Hence, we deemed it preferable to elaborate the following within this framework for consistency with other studies and mainstream methodological literature.

2.2 Cox model using log-ratio coordinates

As the ilr-coordinates z_i, $i = 1, \dots, D-1$, obtained from the original composition using [3] are real variables, they can be incorporated into the hazard function of the standard Cox regression model [2] in the usual manner. In practice however, the hazard function is likely to include additional covariates, e.g. confounding variables. Thus, we extend the Cox model [2] to include both the ilr coordinates and the additional covariates, so that

$$y = \ln\left(\frac{h(t;\boldsymbol{w})}{h_0(t)}\right) = \sum_{j=1}^{D-1}\gamma_j z_j + \boldsymbol{\beta}^T\boldsymbol{v} \ = \boldsymbol{\gamma}^T\boldsymbol{z} + \boldsymbol{\beta}^T\boldsymbol{v}, \qquad [6]$$

where the vectors $\boldsymbol{z}$ and $\boldsymbol{v}$ account for the ilr-coordinates and any other covariates respectively to form the entire vector of explanatory variables $\boldsymbol{w} = (\boldsymbol{z}^T, \boldsymbol{v}^T)^T$ and $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are the corresponding vectors of regression coefficients forming $\boldsymbol{\alpha} = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}^T)^T$. These coefficients can be fitted in the usual manner by maximizing the partial likelihood function

$$L_1(\boldsymbol{\beta},\boldsymbol{\gamma}) = \prod_{j=1}^{K}\left\{\sum_{\left(k_1,\cdots,k_{d_j}\right)=\left(1,\cdots,d_j\right)}\prod_{i=1}^{d_j}\left(\frac{\exp(\boldsymbol{\gamma}^T\boldsymbol{z}_{(j)}^{(k_i)}+\boldsymbol{\beta}^T\boldsymbol{v}_{(j)}^{(k_i)})}{\sum_{l\epsilon R_j}\exp(\boldsymbol{\gamma}^T\boldsymbol{z}_l+\boldsymbol{\beta}^T\boldsymbol{v}_l)-\sum_{s=i}^{d_j}\exp(\boldsymbol{\gamma}^T\boldsymbol{z}_{(j)}^{(s)}+\boldsymbol{\beta}^T\boldsymbol{v}_{(j)}^{(s)})}\right)\right\}, \qquad [7]$$

where K is the number of distinct event times and d_j is the number of individuals with events at time t_j, with $(t_j)_{\ j\in\{1,\cdots,K\}}$ being the distinct event times, and R_j being the set of individuals exposed to risk just prior to event time t_j. The terms $\boldsymbol{z}_l$ and $\boldsymbol{v}_l$ denote respectively the sets of ilr-coordinates and non-compositional covariates for individual l in the set R_j, $\boldsymbol{z}_{(j)}^{(k_i)}$ and $\boldsymbol{v}_{(j)}^{(k_i)}$ denote respectively the sets of ilr-coordinates and non-compositional covariates for individual k_i in the set of individuals with events at time t_j. The subscript in L_1 indicates this is the partial likelihood, to be distinguished from the normal likelihood function L. The expression above allows for the possibility of tied event times, and involves summation over all possible rankings of concurrent events. In practice, the expression is often simplified using the so-called Breslow or Efron approximations (30,31). Note that there is no closed-form solution of this maximum likelihood problem available. Hence, numerical routines are used by most statistical packages for obtaining approximate estimates of the model coefficients.

If the ilr-coordinates chosen are rotated using the rotation matrix $\boldsymbol{H}$, the corresponding coefficients in the partial likelihood components can simply be obtained as

$$\boldsymbol{\gamma}^* = \boldsymbol{H}\boldsymbol{\gamma}. \qquad [8]$$

Model selection for Cox regression is typically conducted using a modified Akaike Information Criterion (AIC) which replaces the likelihood in the standard expression with the partial likelihood above (or some approximation thereof). Again, once the composition has been mapped into real space expressed through ilr-coordinates, there are no obstacles to continue using standard model selection and validation tools.

2.3 Hypothesis testing

When fitting a Cox regression model, it is important to assess the appropriateness of the underlying proportional hazards assumption. Standard statistical tests of this assumption, such as the test of Harrel and Lee (32), remain appropriate, however a new complication arises. These tests, and also

graphical methods such as observed vs. predicted survivor curves and estimated log-log survival curves (33), are applied to individual explanatory variables, and this clashes with the essentially inseparable multivariate nature of compositional variables. For this reason, it is appropriate to consider the validity of the proportional hazard assumption with respect to the whole composition by adapting global tests such as Grambsch and Therneau's non-proportionality test statistic based on Schoenfeld residuals (34).

The Grambsch-Therneau test statistic is defined with reference to a specific hypotheses of time dependence, where a regression coefficient α_j, which is assumed to be constant for the usual Cox model, is hypothesized to be a function of time of the form

$$\alpha_j(t) = \alpha_j + \theta_j g_j(t), \quad [9]$$

where $g_j(t)$ is a deterministic function of t. Common choices are $g_j(t) = t$ and $g_j(t) = \log(t)$, however we have generally used a scaling to time based on the Kaplan-Meier curve:

$$g(t_k) = 1 - S_{KM}(t_k), \quad [10]$$

where

$$S_{KM}(t) = \prod_{k:\, t_k \le t} \left(1 - {}^{d_k}/_{n_k}\right), \quad [11]$$

with d_k being the number of events at time t_k, and n_k being the number of individuals at risk at time t_k (prior to the d_k deaths). This latter choice reduces the sensitivity of the test to a small number of outliers, which can be advantageous in some circumstances (35).

The time dependence is defined for all J covariates, and can be expressed as a diagonal matrix $\boldsymbol{G}$:

$$\boldsymbol{G} = \begin{bmatrix} g_1(t) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & g_J(t) \end{bmatrix} = \begin{bmatrix} \boldsymbol{G}_c & \mathbf{0} \\ \mathbf{0} & \boldsymbol{G}_n \end{bmatrix}, \quad [12]$$

where $\boldsymbol{G}_c$ is the $(D-1) \times (D-1)$ diagonal matrix in respect to the ilr-coordinates, and $\boldsymbol{G}_n$ is the $(J-D+1) \times (J-D+1)$ diagonal matrix in respect to the non-compositional covariates. If the intention is to test the proportional hazards in respect to the composition alone, then all entries of $\boldsymbol{G}_n$ should be set to zero.

The test statistic can now be calculated as a sum over the distinct times of the K events $\{t_k\}_{k=1,..,K}$ as

$$T(\boldsymbol{G}) = \left(\sum_{k=1}^{K} \boldsymbol{G}_k\, \hat{\boldsymbol{r}}_k\right)^T \boldsymbol{D}^{-1} \left(\sum_{k=1}^{K} \boldsymbol{G}_k\, \hat{\boldsymbol{r}}_k\right), \quad [13]$$

where

$$\hat{\boldsymbol{r}}_k = \begin{pmatrix}\boldsymbol{z}_k \\ \boldsymbol{v}_k\end{pmatrix} - \widetilde{\begin{pmatrix}\boldsymbol{z}_k \\ \boldsymbol{v}_k\end{pmatrix}} = \begin{pmatrix}\boldsymbol{z}_k \\ \boldsymbol{v}_k\end{pmatrix} - \frac{\sum_{l\in R_l}\binom{\boldsymbol{z}_l}{\boldsymbol{v}_l} exp(\hat{\boldsymbol{\gamma}}^T\boldsymbol{z}_l + \hat{\boldsymbol{\beta}}^T\boldsymbol{v}_l)}{\sum_{l\in R_l} exp(\hat{\boldsymbol{\gamma}}^T\boldsymbol{z}_l + \hat{\boldsymbol{\beta}}^T\boldsymbol{v}_l)} \quad [14]$$

are the Schoenfeld residuals at time t_k, defined in the usual manner in terms of the difference between the covariates of the individual with an event at time t_k and the weighted average of covariates over all individuals exposed to risk at time t_k (but we have made the distinction between the ilr-coordinates $\boldsymbol{z}_k$ and the non-compositional covariates $\boldsymbol{v}_k$ explicit),

$$\boldsymbol{G}_k = \begin{bmatrix} g_1(t_k) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & g_J(t_k) \end{bmatrix} \quad [15]$$

is the diagonal matrix $\boldsymbol{G}$ calculated at time t_k, and

$$\boldsymbol{D} = \sum_{k=1}^{K} \boldsymbol{G}_k\, \widehat{\boldsymbol{V}}_k \boldsymbol{G}_k{}^T - \left(\sum_{k=1}^{K} \boldsymbol{G}_k\, \widehat{\boldsymbol{V}}_k\right) \widehat{\boldsymbol{V}}_k{}^{-1} \left(\sum_{k=1}^{K} \boldsymbol{G}_k\, \widehat{\boldsymbol{V}}_k\right)^T, \quad [16]$$

based on

$$\widehat{\boldsymbol{V}}_k = \frac{\boldsymbol{S}^{(2)}(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}},t_k)}{S^{(0)}(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}},t_k)} - \left\{\frac{\boldsymbol{S}^{(1)}(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}},t_k)}{S^{(0)}(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}},t_k)}\right\} \left\{\frac{\boldsymbol{S}^{(1)}(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}},t_k)}{S^{(0)}(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}},t_k)}\right\}^T, \quad [17]$$

where

$$S^{(0)}\left(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}},t_k\right) = \sum_{i=1}^{n} \delta_i(t_k) \exp(\hat{\boldsymbol{\gamma}}^T\boldsymbol{z}_i + \hat{\boldsymbol{\beta}}^T\boldsymbol{v}_i)\,), \quad [18]$$

$$\boldsymbol{S}^{(1)}\left(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}},t_k\right) = \sum_{i=1}^{n} \delta_i(t_k) \exp(\hat{\boldsymbol{\gamma}}^T\boldsymbol{z}_i + \hat{\boldsymbol{\beta}}^T\boldsymbol{v}_i)\, \begin{pmatrix}\boldsymbol{z}_i \\ \boldsymbol{v}_i\end{pmatrix} \text{ and} \quad [19]$$

$$\boldsymbol{S}^{(2)}\left(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}},t_k\right) = \sum_{i=1}^{n} \delta_i(t_k) \exp(\hat{\boldsymbol{\gamma}}^T\boldsymbol{z}_i + \hat{\boldsymbol{\beta}}^T\boldsymbol{v}_i)\, \begin{pmatrix}\boldsymbol{z}_i \\ \boldsymbol{v}_i\end{pmatrix}\begin{pmatrix}\boldsymbol{z}_i \\ \boldsymbol{v}_i\end{pmatrix}^T. \quad [20]$$

This test statistic is defined for the complete set of explanatory variables, but it can be applied separately to a subset of them. Thus, we adapt it to deal with the compositional variables only. Importantly, it can be checked that this test is invariant to rotations of the ilr-coordinates, provided the hypothesized time dependence is identical for each ilr-coordinate. That is,

$$\boldsymbol{G}_c = g(t)\, \boldsymbol{I}_{D-1} \quad [21]$$

as can be seen by substituting

$$\boldsymbol{\gamma}^* = \boldsymbol{H}\boldsymbol{\gamma} \text{ and } \boldsymbol{z}_l^* = \boldsymbol{H}\boldsymbol{z}_l \quad [22]$$

into the Eq. [13]. The formal proof is included as Appendix A.

Moreover, we consider a test of the significance of the association between the hazard rate and the composition as a whole. The natural approach is to adapt the likelihood ratio test using the partial likelihood from Eq. [7] based on the statistic

$$LRT = 2\,[\ln(L_1^1) - \ln(L_1^0)], \quad [23]$$

where L_1^0 is the partial likelihood for the model omitting the ilr-coordinates, and L_1^1 is the partial likelihood for the complete model including the ilr-coordinates. Under the null hypothesis, this test statistic is asymptotically distributed as a chi-square distribution with $D-1$ degrees of freedom (assuming a D-part composition). Note that this test is invariant to rotations of the ilr-coordinates as can be seen by substituting Eq. [22] into Eq. [23].

3. Hazard ratios with compositional covariates

Hazard ratios are commonly used in the interpretation of the output from a fitted Cox regression model. To use them it is necessary to define a baseline set of explanatory variables $\boldsymbol{w}_0 = (\boldsymbol{z}_0{}^T, \boldsymbol{v}_0{}^T)^T$, for the comparison, where $\boldsymbol{z}_0$ represents the ilr-coordinates at the baseline composition, and $\boldsymbol{v}_0$ represents the non-compositional baseline covariates. The hazard ratio is then defined as

$$HR = \frac{h(t;\boldsymbol{w})}{h(t;\boldsymbol{w}_0)} = \frac{\exp(\boldsymbol{\gamma}^T\boldsymbol{z} + \boldsymbol{\beta}^T\boldsymbol{v})}{\exp(\boldsymbol{\gamma}^T\boldsymbol{z}_0 + \boldsymbol{\beta}^T\boldsymbol{v}_0)}.$$

This presents a complication for compositional variables as zero-values for the ilr-coordinates (corresponding to equal allocations to each component in the raw compositional data) may not give a baseline that is meaningful in terms of a real-world problem. The most natural option would be a composition corresponding to the mean ilr-coordinates, which conveniently corresponds by

isometry with the compositional centre in the raw compositional data set (rescaled appropriately) (11). However, the scientific question being investigated can also inform this choice.

Note that using Eq. [3], the exponential part in the hazard function $h(t;\boldsymbol{w})$ can be broken down into a product of exponents as follows:

$$\exp\left(\sum_{i=1}^{D-1}\gamma_i z_i + \boldsymbol{\beta}^T\boldsymbol{v}\right)$$

$$= \exp(\boldsymbol{\beta}^T\boldsymbol{v})\prod_{i=1}^{D-1}\exp\left(\gamma_i \,.\, \mathrm{z}_i\right)$$

$$= \exp(\boldsymbol{\beta}^T\boldsymbol{v})\prod_{i=1}^{D-1}\exp\left(a_i \cdot \ln\left[\frac{\left(\prod_{k=1}^{r_i} x_{i_k}^{+}\right)^{\frac{1}{r_i}}}{\left(\prod_{l=1}^{s_i} x_{i_l}^{-}\right)^{\frac{1}{s_i}}}\right]\right)$$

$$= \exp(\boldsymbol{\beta}^T\boldsymbol{v})\prod_{i=1}^{D-1}\left[\frac{\left(\prod_{k=1}^{r_i} x_{i_k}^{+}\right)^{\frac{1}{r_i}}}{\left(\prod_{l=1}^{s_i} x_{i_l}^{-}\right)^{\frac{1}{s_i}}}\right]^{a_i}$$

$$= \exp(\boldsymbol{\beta}^T\boldsymbol{v})\prod_{j=1}^{D}\left(x_j\right)^{c_j}, \qquad [24]$$

where

$$a_i = \sqrt{\frac{r_i \cdot s_i}{r_i + s_i}}\,\gamma_i\,, \qquad [25]$$

$$c_j = \sum_{i=1}^{D-1}\left(\frac{\delta_{ij}^{+}}{r_i} - \frac{\delta_{ij}^{-}}{s_i}\right)a_i \qquad [26]$$

based on

$$\delta_{ij}^{+} = \begin{cases} 1 & \text{when } x_j \text{ appears in numerator of balance } z_i \\ 0 & \text{otherwise} \end{cases}$$

$$\text{and } \delta_{ij}^{-} = \begin{cases} 1 & \text{when } x_j \text{ appears in denominator of balance } z_i \\ 0 & \text{otherwise} \end{cases}$$

Essentially, the use of the Cox regression model in combination with an ilr representation of compositional covariates enables a simple expression for the hazard function in terms of the product of the components of the composition raised to indices determined by the fitted regression coefficients. When the objective of the analysis is to determine the association of the compositional variables with the survival outcome, then the expression for the hazard ratio (HR) between a

composition $\boldsymbol{x}$ and the baseline composition $\boldsymbol{x}_0$ can be expressed entirely in terms of the compositional variables as

$$HR = \prod_{j=1}^{D} \left(\frac{x_j}{x_{0j}}\right)^{c_j}, \quad [27]$$

with the non-compositional covariates vanishing from the hazard ratio entirely in correspondence with the features of the proportional hazards model.

At first sight the coefficients $\left(c_j\right)_{j=1,\ldots,D}$ seem to provide a means to isolate the influence of a single component whilst allowing for the adequate compositional treatment of the data. However, these parameters are constrained, and it is not actually possible to alter a single component in isolation. Confidence intervals for the parameters c_j are simple to obtain based on the end points of the confidence interval for the ordinary Cox model parameters. Thus, it is straightforward to determine if the interval includes 0, but this may not be an appropriate null hypothesis. Reallocating from a component with a positive coefficient to a component with a "zero" coefficient will be predicted to reduce the hazard function, therefore the zero coefficient does not imply the component has no effect as it does for conventional variables. Comparing the confidence intervals of the parameters c_j pairwise to determine which do not overlap is also inappropriate as the treatment of the compositional variables implies that changes occur on a proportional scale. Therefore, the effect of a fixed reallocation (in the original units the compositions were measured in) between two components, say x_1 and x_2, will be based on this fixed reallocation as proportions of the original values of the two components, rather than just the size of the change in the original units. For example, in time-use compositions, reallocating 30 minutes between two behaviours will have an effect that depends on the original allocation to the two behaviours, not just the 30 minutes. As a result, it is not possible to define a single number quantifying the relationship to the original components. The nearest we come to this is by rewriting the equation in the form

$$HR = A\,(x_1)^{c_1}\,(x_2)^{c_2}, \quad [28]$$

where

$$A = \left(\frac{1}{x_{01}}\right)^{c_1}\left(\frac{1}{x_{02}}\right)^{c_2}\prod_{j=3}^{D}\left(\frac{x_j}{x_{0j}}\right)^{c_j}. \quad [29]$$

This will remain constant if we consider only exchanges between two parts at a time (without loss of generality). It is therefore possible to produce two dimensional graphs of the hazard ratio against (x_1, x_2) and so illustrate the consequences of reallocating e.g. time between pairings of physical activity behaviour categories. Such an approach may be of limited value for large D though. However, for low-moderate dimension problems, this provides a means to understand the relationship between the outcome and reallocations between the different parts of the composition. The approach could be generalised to exchanges between more than two parts, considering groups of parts, but the proportions in which these are reallocated must then be specified to avoid ambiguity. Restricting ourselves to two parts gives a clear and simple message that can be readily understood by practitioners in the relevant field. It would be also possible to consider the association of reallocating between particular parts by constructing regression models from simple log-ratios representing that trade-off (see e.g. (36) in a high-dimensional context).

Finally, note that when working with compositional covariates it would be acceptable to centre their ilr-coordinates to use $\tilde{z}_i = z_i - \bar{z}_i$, as commonly done with ordinary covariates to make survival and hazard functions relative to the mean rather than relative to the minimum, which is usually the most meaningful comparison. However, further transforming the ilr-coordinates into z-scores by dividing by the corresponding standard deviations is not advisable. The ilr-coordinates are already on the same scale and their relative variability provides relevant information which would be lost with that operation.

4. Application to physical activity research: association between mortality and physical activity patterns from the NHANES survey

To illustrate the use of the compositional Cox regression model developed in the previous sections, we set out a case study from physical activity research using data from the NHANES 2005-06

database. NHANES is a bi-annual health survey conducted by the Centre for Disease Control on a representative sample of the population of the United States of America. The 2005-06 wave included an assessment of physical activity behaviour and sleep. Further details of the NHANES survey can be found on the Centre for Disease Control website. All analyses were conducted on the R system for statistical computing (37), using the *survival* package (38) for the Cox regression fitting, the *ggfortify* package (39) to produce the Kaplan-Meier survival curves and the *ggtern* package (40) to create the ternary plots. This study involved analysis of publicly available secondary data only. The original study was approved by the ethics committee of the Centers for Disease Control and Prevention (CDC) and all participants gave informed consent. NHANES operates under the approval of the National Center for Health Statistics Research Ethics Review Board, Protocols #98-12, #2005-06, and #2011-17.

Here, we consider the link between all-cause mortality and the time-use composition of the day (after allowing for other relevant covariates). The time-use composition consisted of sleep (S), sedentary behaviour (SB), light intensity physical activity (LIPA) and moderate to vigorous physical activity (MVPA), hence a composition $\boldsymbol{x} = (\mathrm{S}, \mathrm{SB}, \mathrm{LIPA}, \mathrm{MVPA}) \in S^4$, expressed in percentages relative to the 24-hour day. SB, LIPA and MVPA were computed from accelerometry data using standard processing methods for NHANES, whereas sleep time was self-reported (41). We restricted the analysis to individuals aged between 50-79 years and removed one accidental death and one individual with invalid accelerometer data. This is in line with previous work done on NHANES data, and helps to avoid violations of the proportional hazards assumption (42).

After removing these individuals, there were 1,196 records, including 114 deaths. The total exposure of the data was 82,165 person months. Five records had zero-valued MVPA. These zeroes were imputed using the log-ratio EM algorithm (43). The resulting mean composition (computed as the vector of geometric mean times spent on each behaviour closed to 100%) was $g(\boldsymbol{x}) = (29.1, 44.4, 25.9, 0.7)\%$. Kaplan-Meier survival curves by gender and age group are shown in Figure 1.

[Include Figure1.tif]

Figure 1: Kaplan-Meier survival curves for subset of NHANES 05-06 data stratified by gender and age tertile (+ symbols indicate an observation ceasing on an individual rather than a death).

A set of three ilr-coordinates $\{z_1, z_2, z_3\}$ is obtained from this 4-part composition. The set of ilr-coordinates $\boldsymbol{z}^1$ below was constructed by SBP so that the first coordinate represented a contrast between sleep time (the part x_1 in $\boldsymbol{x}$ above used as reference; see Eq. [5]) and waking day over the 24-hour day. The second ilr-coordinate in this same set broadly distinguished between sedentary and active behaviour. Finally, the third one compared the different intensities of physical activity. Namely, the corresponding formal expressions were

$$z_1^1 = \sqrt{\frac{3}{4}} \ln \frac{S}{(MVPA \cdot LIPA \cdot SB)^{1/3}} \quad , \qquad [30]$$

$$z_2^1 = \sqrt{\frac{2}{3}} \ln \frac{SB}{(MVPA \cdot LIPA)^{1/2}} \text{ and} \qquad [31]$$

$$z_3^1 = \sqrt{\frac{1}{2}} \ln \frac{LIPA}{MVPA}. \qquad [32]$$

These ilr-coordinates were then incorporated into the Cox model [6] to produce the following Model 1:

$$y = \gamma_1 z_1^1 + \gamma_2 z_2^1 + \gamma_3 z_3^1 + \boldsymbol{\beta}^T \boldsymbol{v}. \qquad [33]$$

The covariates in $\boldsymbol{v}$ were age, sex, smoking status, alcohol consumption, and energy intake. To explore which relative time allocations drive any association, we considered alternative ilr-coordinate triplet sets $\boldsymbol{z}^2, \boldsymbol{z}^3$ and $\boldsymbol{z}^4$ by orthogonal rotation:

$$\boldsymbol{z}^2 = \left\{ \sqrt{\frac{3}{4}} \ln \frac{SB}{(S \cdot LIPA \cdot MVPA)^{1/3}}, \sqrt{\frac{2}{3}} \ln \frac{S}{(LIPA \cdot MVPA)^{1/2}}, \sqrt{\frac{1}{2}} \ln \frac{LIPA}{MVPA} \right\} \qquad [34]$$

$$\boldsymbol{z}^3 = \left\{ \sqrt{\frac{3}{4}} \ln \frac{LIPA}{(S \cdot SB \cdot MVPA)^{1/3}}, \sqrt{\frac{2}{3}} \ln \frac{S}{(SB \cdot MVPA)^{1/2}}, \sqrt{\frac{1}{2}} \ln \frac{SB}{MVPA} \right\} \text{and} \qquad [35]$$

$$\boldsymbol{z}^4 = \left\{ \sqrt{\frac{3}{4}} \ln \frac{MVPA}{(S \cdot SB \cdot LIPA)^{1/3}}, \sqrt{\frac{2}{3}} \ln \frac{S}{(SB \cdot LIPA)^{1/2}}, \sqrt{\frac{1}{2}} \ln \frac{SB}{LIPA} \right\}. \qquad [36]$$

Each one of these triplets isolated the relative importance of SB, LIPA, and MVPA against the other behaviors in the first ilr-coordinate respectively. These ilr-coordinates can then be incorporated into Eq. [6] in the usual manner, given rise to an equivalent Model 2 in terms of $\boldsymbol{z}^2$:

$$y = \gamma_1 z_1^2 + \gamma_2 z_2^2 + \gamma_3 z_3^2 + \boldsymbol{\beta}^T \boldsymbol{v}, \quad [37]$$

and similarly to Models 3 and 4 with respect to $\boldsymbol{z}^3$ and $\boldsymbol{z}^4$.

In the following we seek to test the proportional hazards assumption underlying the model. The Gramsch-Therneau's test (Eq. [13]) requires a hypothesized time dependence. We have used a scaling to time based on the Kaplan-Meier curve, set out in Eq. [9-11] (this is the default time dependence used in the *survival* package in R). Based on this hypothesis, we obtained a test p-value equal to 0.106, indicating no support for rejecting the proportional hazards assumption at the usual 5% significance level.

The statistical significance of the association between mortality and time-use composition as a whole, after allowing for other covariates, was assessed by comparing the proposed model (either Model 1 or Model 2 above) with a baseline model based on the non-compositional covariates only using the likelihood ratio test. The obtained p-value of 0.02 indicated a statistically significant difference between them. Therefore, we found support for a statistically significant association between mortality hazard and the physical activity composition of the 24-hour day, beyond that attributable to other covariates. The fitted coefficients associated with the compositional Cox regression model are summarized in Table 1 using all four model formulations, alongside the p-value for the overall composition. Note that the p-value equal to 0.02 for the overall effect of the composition is the same for all model formulations as expected.

Model	Ilr-coord.	exp(γ)	Lower Bound	Upper Bound	p-value*
1	z_1^1	1.0450	0.5277	2.0692	0.90
	z_2^1	1.4745	0.8878	2.4490	0.13
	z_3^1	1.0339	0.6723	1.5900	0.88
	Overall	-	-	-	0.02

2	z_1^2	1.4211	0.7642	2.6428	0.27
	z_2^2	1.1864	0.6627	2.1241	0.57
	z_3^2	1.0339	0.6723	1.5900	0.88
	Overall	-	-	-	0.02
3	z_1^3	0.8432	0.4798	1.4819	0.55
	z_2^3	0.9865	0.5038	1.9317	0.97
	z_3^3	1.4233	0.9791	2.0688	0.06
	Overall	-	-	-	0.02
4	z_1^4	0.7986	0.6490	0.9826	0.03
	z_2^4	0.9677	0.4681	2.0006	0.93
	z_3^4	1.3766	0.7691	2.4638	0.28
	Overall	-	-	-	0.02

*p-values for individual ilr-coordinates are based on Wald tests and overall p-values are based on (Partial) likelihood ratio tests.

Table 1: Cox regression coefficients and 95% confidence limits of ilr-coordinates under different model formulations.

In respect of the individual terms, only the term z_1^4 was statistically significant (p = 0.03), thus pointing at the proportion of time allocated to MVPA, relative to the geometric mean of the other behaviors considered, as the main behavior allocation responsible of the beneficial association between mortality and physical activity. This can also be seen from the upper and lower bounds on the 95% confidence interval which are both less than 1. We note the conclusions remain unchanged (although the results are not identical) if sleep is omitted from the analysis and we consider only the subcomposition of the waking day. If we consider the second coordinate in model 4, z_2^4, we can see it is not statistically significant (p-value = 0.98) and it would be reasonable to consider eliminating this coordinate from our model subject to the same considerations we would apply for a conventional (non-compositional) covariate, indicating the proportion of time allocated

to sleep (after allowing for the proportion of time allocated to MVPA) is not associated with mortality. In fact, we can go further, and consider eliminating z_3^4 after eliminating z_2^4 (p-value = 0.28), suggesting this dataset only supports an association between mortality and the proportion of MVPA relative to other behaviors. However, for the purposes of illustration, we will continue to work with the full model in presenting results.

Alternatively, it is reasonable to consider an alternative basis for the subcomposition (LIPA, SB, Sleep) by rotating the ilr-coordinates z_2^4 and z_3^4 giving a new basis, e.g.

$$\tilde{\boldsymbol{z}}^4 = \left\{ \sqrt{\frac{3}{4}} \ln \frac{MVPA}{(S \cdot SB \cdot LIPA)^{1/3}}, \sqrt{\frac{2}{3}} \ln \frac{LIPA}{(SB \cdot S)^{1/2}}, \sqrt{\frac{1}{2}} \ln \frac{SB}{S} \right\}$$

and due to the orthonormality of these coordinates, the coefficient and p-value of the first coordinate z_1^4 are invariant to this transform. If one considers the physical behavior types to represent different intensities of activity on a continuous scale from MVPA (highest intensity activity) to LIPA (low intensity activity) and then to SB and Sleep (non-active behaviors) then this alternative basis can be regarded broadly as a series of hypotheses testing the presence of an association between mortality and decreasing intensities of activity. Including Sleep in such a continuum is not wholly satisfactory, however it illustrates how one can use the construction of balances in testing different hypotheses. An alternative approach would be to use the alr transformed variables to construct hypotheses (see e.g. (14) in the context of linear regression).

A key feature of Cox regression is the ability to express the survival probability relative to some baseline as a hazard ratio. Where the intention is to determine the association of mortality and the composition of physical activity, we would typically hold other covariates fixed, and so the hazard ratio becomes

$$\mathrm{HR} = \left(\frac{S}{S_0}\right)^{c_1} \left(\frac{SB}{SB_0}\right)^{c_2} \left(\frac{LIPA}{LIPA_0}\right)^{c_3} \left(\frac{MVPA}{MVPA_0}\right)^{c_4} \quad [38]$$

where $c_1 = 0.044$, $c_2 = 0.374$, $c_3 = -0.175$, $c_4 = -0.242$, based on Eq. [24-26].

The size and sign of these coefficients could naively be considered to give an indication of the strength of the association between mortality and each behaviour, and whether the behaviour is beneficial (negative coefficient) or detrimental (positive coefficient). However, it should be recalled that these parameters are constrained, and as noted in the previous section, it is not possible to alter a single component in isolation, meaning the individual values are not really meaningful. The values of these coefficients can be considered pairwise to some benefit. For example, we note $c_2 > c_1$ indicating that reallocating time from SB to Sleep is predicted to reduce an individual's mortality risk.

Rewriting this equation as

$$\mathrm{HR} = \left(1 + \frac{S - S_0}{S_0}\right)^{c_1} \left(1 + \frac{SB - SB_0}{SB_0}\right)^{c_2} \left(1 + \frac{LIPA - LIPA_0}{LIPA_0}\right)^{c_3} \left(1 + \frac{MVPA - MVPA_0}{MVPA_0}\right)^{c_4}$$

provides some further insight. In particular, the association of mortality rate with reallocating time to or from a given behaviour type will depend on the current level of that behaviour type. Therefore, comparing the values of the coefficients directly is informative if a reference composition has been specified. A comprehensive representation of the effects of different reallocations on mortality risk is found in Figure 3. An alternative to investigate the association of a particular exchange between behaviours and HR would be to define a set of ilr-coordinates through SBP which includes a particular exchange of interest in the form of a balance (e.g. SB to Sleep). However, this would provide the same results shown in Figure 3, as this would be an orthogonal rotation of the ilr bases used for Model 1-4. Note that the idea of focusing only on a simple regression on that balance of interest would ignore the potential influence on HR of exchanges between behaviours as represented in the other balances in the ilr-coordinate set.

The choice of reference composition is the practitioner's decision. Defaulting to ilr-coordinates equal to zero (i.e. an even distribution of time across all four components) would be unrealistic (e.g. it would mean considering 25% MVPA = 6 hours/day). Using zero activity as reference composition would not be feasible as that would not actually define a composition and log-ratios cannot be

computed. We instead used the geometric mean physical activity composition as a sensible reference. However, note that any other could be used. For example, one representing an individual complying with some set of national guidelines on physical activity.

We used a ternary heatmap to represent the outcome of our model. As it is a 4-part composition but only three can be represented in a ternary plot, we plotted the hazard ratio (relative to the average composition baseline) against the four possible subcompositions. In Figure 2(a) we omit Sleep, and illustrate the hazard ratio against the subcomposition (SB, LIPA, MVPA) whilst holding Sleep fixed. In Figure 2(b) we hold SB fixed and consider the hazard ratio against (Sleep, LIPA, MVPA). In Figure 2(c) we hold LIPA fixed and consider the hazard ratio against (Sleep, SB, MVPA). In Figure 2(d) we hold MVPA fixed and consider the hazard ratio against (Sleep, SB, LIPA). For a given model, the allocation of time to a fixed component does not affect the relationship between hazard ratio and the subcomposition in accordance with the subcompositional coherence property of log-ratio analysis. However, if a different model is fitted based only on the subcomposition there will be small differences in the hazard ratio as information from one part is not available. The blue point indicates the reference composition in each ternary plot.

Alternatively, the impact of reallocations between two particular behaviours can be illustrated graphically in a manner similar to the isotemporal substitution analysis commonly used in physical activity research (44). Figure 3(a) shows the hazard ratio against the time allocated to MVPA assuming the only permitted compositions are fixed amounts of time reallocated between MVPA and another component of the composition. For example, the green line indicates the effect of reallocating time between MVPA and SB, holding Sleep and LIPA fixed at the compositional average. Similarly, Figure 3(b) shows the hazard ratio against time allocated to LIPA, Figure 3(c) shows the hazard ratio against time allocated to SB, and Figure 3(d) shows the hazard ratio against time allocated to Sleep. In each case, time is reallocated between the component displayed on the x-axis and the component indicated by the colour of the line. Figure 3(a) clearly illustrates that reallocating time to MVPA from the other three behaviours is associated with lower mortality but that the

beneficial association weakens as MVPA increases. Figure 3(b) confirms that reallocating time from MVPA to LIPA increases the mortality rate, but reallocating time from Sleep or SB to LIPA is associated with lower mortality, however the association is less strong than for MVPA. Figures 3(c) and 3(d) can be read in the same manner. It is worth remarking that the time allocated to each behaviour is constrained to be positive, hence one cannot reallocate time between two behaviours without limit and where a line stops mid-graph this indicates the replaced behaviour has reached zero.

4.1 Mortality and balance between active and non-active behaviours

As noted in Section 2.1.3, the parts of the composition can be meaningfully partitioned into two or more subsets of practical relevance. For example, we can define an ilr-representation $\boldsymbol{z}^5$ using SBP (see supplementary materials) which includes the compositional balance between active behaviours $(\mathrm{LIPA}, \mathrm{MVPA})$ and non-active behaviours $(\mathrm{S}, \mathrm{SB})$:

$$\boldsymbol{z}^5 = \left\{ \ln \frac{(MVPA \cdot LIPA)^{1/2}}{(S \cdot SB)^{1/2}}, \sqrt{\frac{1}{2}} \ln \frac{MVPA}{LIPA}, \sqrt{\frac{1}{2}} \ln \frac{S}{SB} \right\}.$$

The Cox regression model was fitted to these ilr-coordinates as in the previous section, and we focused on the change in the hazard function associated with an increase of 1 unit in the first coordinate through $\exp(\gamma_1)$ (value: 0.71; 95% confidence interval: (0.47, 1.07); p-value: 0.104). This indicated that increases in the (geometric) average active behaviour were associated with reduced mortality. However, unlike the case of Model 4 above where the balance between MVPA and the other behaviours was considered, this relationship is not strictly statistically significant at the usual 5% significance level. This is in agreement with previous evidence that the link between physical activity and mortality is mostly driven by the relative allocation to MVPA (45), whereas the association with of LIPA is more contentious (46).

5. Discussion

In this work we have shown how to extend survival data analysis based on the Cox regression model to deal with compositional covariables using the log-ratio methodology. It is well-established that fitting raw compositional variables in ordinary regression analysis introduces both technical and interpretability issues due to their failure to account for their specific nature induced by scale invariance. Compositional data analysis is being increasingly and successfully used in cross-sectional studies, however to date it has not been introduced in survival data analysis to the best of our knowledge.

Survival analysis based on compositional data has been applied in the past that do not adopt this log-ratio methodology, however inadequacies in the conventional approach as stressed above can lead to misleading estimates, as some of the effects attributed to a component (in the absolute sense) will in fact arise from the displacement of other components, and this may give qualitative conclusions that are misleading as well, particularly for smaller marginal effects, such as the associations between mortality and levels of light intensity physical activity.

The resolution of the formal issues around compositional data analysis requires some changes in how the results are interpreted which can be initially perceived as less intuitive and more challenging, as it deviates from the usual variable-by-variable analysis. In our context of application, the use of log-ratio coordinates prompts the discussion in terms of changes in behaviours which are always relative to each other. This in fact aligns with the intuitive idea that changes in time devoted to an activity are necessarily linked to changes in opposite direction in some other activity or activities, and this exchange may have implications to mortality risk as investigated here. As stressed above, the choice of specific log-ratio coordinates can be guided by practical and scientific considerations of the domain of application, so that they are tailored to investigate the most relevant research questions in each case. This will facilitate the communication of conclusions in a way which is meaningful for the target audience. The relative merits of alternative log-ratio

representations are an open topic of debate and are out of the scope of this work. However, it is important to remark that the differences are only in how the information in the composition is represented using log-ratios. Thus, overall estimates, test statistics, performance measures or predictions, including the coefficient of determination, from the Cox regression model will be the same. When working with large compositions, it may also be beneficial to consider the elimination of individual log-ratio coordinates through e.g. stepwise regression via the modified AIC, or similar, once the association with the composition has been established to identify its key drivers. As noted previously, models excluding some log-ratio coordinates are perfectly valid. Carried out in tandem with rotation of coordinates, this can bring significant simplifications of the final model, and make clearer which of the internal dynamics of the composition are driving the outcome.

Finally, note that working with log-ratios means that zero values in the data set are problematic. In most real-world situations these values are related to under-reporting, existence of technical detection limits or rounding-off errors. A number of compositional statistical methods have been proposed to deal with them in a sound way as part of the data pre-processing stage (e.g. (43,47–49)). Alternatively, in some cases a meaningful amalgamation of parts into a smaller composition can be a sensible pragmatic choice to address the problem in a real-world setting.

Acknowledgements

D. M. and J. P.-A. have been supported by the Scottish Government's Rural and Environment Science and Analytical Services Division (RESAS). J. P.-A. has also been supported by the Spanish Ministry of Economy and Competitiveness under the project CODA-RETOS MTM2015-65016-C2-1(2)-R. K. H. has been supported by a research grant by the Czech Science Foundation under reg. no. 18-09188S and the grant COST Action CRoNoS IC1408.

The authors declare no conflict of interest, financial or otherwise. The results of the present study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation.

The work was conducted at Glasgow Caledonian University and Biomathematics and Statistics Scotland. All persons designated as authors qualify for authorship, and all those who qualify for authorship are listed.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

[Include Figure2.tif]

Figure 2: Heatmap of hazard ratios against different percentage allocations of physical activity time-use, with fixed (a) Sleep = 29.1%, (b) SB = 44.4%, (c) LIPA = 25.9%, and (d) MVPA = 0.7%. The blue point indicates the reference (geometric) mean composition.

[Include Figure3.tif]

Figure 3: Hazard ratios under isotemporal substitution, reallocating time between (a) MVPA, (b) LIPA, (c) SB or (d) Sleep, and the component indicated in the legend, whilst holding the remaining components fixed) with respect to (geometric) mean composition.

Appendix A: Invariance of the Grambsch-Therneau test statistic to ilr-coordinate system rotation

The Gramsch-Therneau test statistic was defined in Eq. [13-20] with reference to hypothesized time-dependence G, defined in Eq. [9,12]. As noted in Section 2 we restrict this statement to uniform time dependence across the components of the composition of the form outlined in Eq. [21]. For convenience, we repeat Eq. [13] and [21] below:

$$T(\boldsymbol{G}) = \left(\sum_{k=1}^{K} \boldsymbol{G}_k\, \hat{\boldsymbol{r}}_k\right)^T \boldsymbol{D}^{-1} \left(\sum_{k=1}^{K} \boldsymbol{G}_k\, \hat{\boldsymbol{r}}_k\right) \text{ and}$$

$$\boldsymbol{G}_c = g(t)\, \boldsymbol{I}_{D-1}.$$

The covariates (including non-compositional covariates) are defined as

$$\boldsymbol{w}_i = \begin{pmatrix} \boldsymbol{z}_i \\ \boldsymbol{v}_i \end{pmatrix} \quad [39]$$

We require to show that the value of the test statistic does not change when it is recalculated for the rotated covariates:

$$\widetilde{\boldsymbol{w}_i} = \widetilde{\boldsymbol{H}} \begin{pmatrix} \boldsymbol{z}_i \\ \boldsymbol{v}_i \end{pmatrix}, \quad [40]$$

where the rotation matrix is defined as follows:

$$\widetilde{\boldsymbol{H}} = \begin{pmatrix} \boldsymbol{H} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix}. \quad [41]$$

Note that only the ilr-coordinates are rotated. For clarity we also re-express Eq. [19-20] in terms of Eq. [39] as

$$\boldsymbol{S}^{(2)}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}, t_k) = \sum_{i=1}^{n} a_i\, \boldsymbol{w}_i \boldsymbol{w}_i^T \text{ and} \quad [42]$$

$$\boldsymbol{S}^{(1)}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}, t_k) \left\{\boldsymbol{S}^{(1)}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}, t_k)\right\}^T = \sum_{i=1}^{n} b_i\, \boldsymbol{w}_i \boldsymbol{w}_i^T. \quad [43]$$

Note that a_i and b_i are trivially invariant to rotation. Hence if we transform the coordinates based on Eq. [40] these become

$$\tilde{\boldsymbol{S}}^{(2)}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}, t_k) = \sum_{i=1}^{n} a_i\, \widetilde{\boldsymbol{H}} \boldsymbol{w}_i \boldsymbol{w}_i^T \widetilde{\boldsymbol{H}^T} = \widetilde{\boldsymbol{H}} \left(\sum_{i=1}^{n} a_i\, \boldsymbol{w}_i \boldsymbol{w}_i^T\right) \widetilde{\boldsymbol{H}^T} \text{ and} \quad [44]$$

$$\tilde{\boldsymbol{S}}^{(1)}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}, t_k) \left\{\tilde{\boldsymbol{S}}^{(1)}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}, t_k)\right\}^T = \sum_{i=1}^{n} b_i\, \widetilde{\boldsymbol{H}} \boldsymbol{w}_i \boldsymbol{w}_i^T \widetilde{\boldsymbol{H}^T} = \widetilde{\boldsymbol{H}} \left(\sum_{i=1}^{n} b_i\, \boldsymbol{w}_i \boldsymbol{w}_i^T\right) \widetilde{\boldsymbol{H}^T}. \quad [45]$$

In addition, $S^{(0)}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}, t_k)$ is invariant to rotations. Therefore, under the transformed coordinates:

$$\widetilde{\boldsymbol{V}_k} = \widetilde{\boldsymbol{H}} \boldsymbol{V}_k \widetilde{\boldsymbol{H}^T}, \quad [46]$$

$$\widetilde{\boldsymbol{D}} = \widetilde{\boldsymbol{H}}\boldsymbol{D}\widetilde{\boldsymbol{H}^T} \quad \text{and} \tag{47}$$

$$\widetilde{\boldsymbol{D}}^{-1} = \widetilde{\boldsymbol{H}}\boldsymbol{D}^{-1}\widetilde{\boldsymbol{H}^T}. \tag{48}$$

Using associativity of matrix multiplication, and that $\boldsymbol{G_k}$ are defined as matrices of the form

$$\boldsymbol{G}_k = \begin{pmatrix} \boldsymbol{g}(t_k)\,\boldsymbol{I}_{D-1} & \boldsymbol{0} \\ \boldsymbol{0} & \widehat{\boldsymbol{G}_p} \end{pmatrix}, \tag{49}$$

where $\widehat{\boldsymbol{G}_p}$ is a diagonal matrix of the usual (non-compositional) form, and so commute with the square matrix $\widetilde{\boldsymbol{H}}$.

In addition, the Schoenfeld residuals are expressed as $\widehat{\tilde{\boldsymbol{r}}_k} = \widetilde{\boldsymbol{H}}\hat{\boldsymbol{r}}_k$ under rotation. Therefore, the test statistic, under rotation, becomes

$\tilde{T}(\boldsymbol{G}) = (\sum_k \boldsymbol{G}_k\,\hat{\boldsymbol{r}}_k)^T \boldsymbol{H}^T \boldsymbol{D}^{-1} \boldsymbol{H} (\sum_k \boldsymbol{G}_k\,\hat{\boldsymbol{r}}_k) = (\sum_k \boldsymbol{G}_k\,\hat{\boldsymbol{r}}_k)^T \boldsymbol{D}^{-1}\,(\sum_k \boldsymbol{G}_k\,\hat{\boldsymbol{r}}_k) = T(\boldsymbol{G})$. ■

Bibliography

1. Collett D. *Modelling survival data in medical research*. 2nd ed. Boca Raton: CRC, 2003, p. 410.
2. Lee ET and Go OT. Survival analysis in public health research. *Ann Rev Public Health* 1997; 18(1): 105–34.
3. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B* 1972; 34: 187–220.
4. Cox DR. Partial Likelihood. *Biometrika* 1975; 62(2): 269–276.
5. Leite MLC. Applying compositional data methodology to nutritional epidemiology. *Stat Methods Med Res* 2016; 25(6): 3057–3065.
6. Mert MC, Filzmoser P, Endel G, et al. Compositional data analysis in epidemiology. *Stat Methods Med Res* 2018;27(6):1878–1891.
7. Tsilimigras MCB and Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann Epidemiol* 2016; 26(5): 330–335.
8. Gloor GB and Reid G. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Can J Microbiol* 2016; 62(8): 692–703.
9. Chastin SFM, Palarea-Albaladejo J, Dontje ML, et al. Combined effects of time spent in physical activity, sedentary behaviors and sleep on obesity and cardio-metabolic health markers: a novel compositional data analysis approach. *PLoS One* 2015; 10(10): e0139984.
10. Dumuid D, Stanford TE, Martín-Fernández JA, et al. Compositional data analysis for physical activity, sedentary time and sleep research. *Stat Methods Med Res* 2017; 27(12): 3726–3738.
11. Scheffé H. Experiments with mixtures. *J R Stat Soc Ser B* 1958; 20(2): 344–360.
12. Cox DR. A note on polynomial response functions for mixtures. *Biometrika* 1971; 58(1): 155.
13. Cornell JA. *Experiments with mixture-designs, models and the analysis of mixture data*. New York: Wiley, 1981.
14. Aitchison J and Bacon-Shone J. Log contrast models for experiments with mixtures. *Biometrika* 1984; 71(2): 323–330.

15. Aitchison J. Logratios and natural laws in compositional data analysis. *Math Geol* 1999; 31(5): 563–580.

16. Pawlowsky-Glahn V, Egozcue JJ and Tolosana-Delgado R. *Modelling and analysis of compositional data*. Chichester: Wiley, 2015, p. 272.

17. Lin W, Shi P, Feng R, et al. Variable selection in regression with compositional covariates. *Biometrika* 2014; 101(4): 785–797.

18. Shi P, Zhang A and Li H. Regression analysis for microbiome compositional data. *Ann Appl Stat* 2016; 10(2): 1019–1040.

19. Müller I, Hron K, Fišerová E, et al. Interpretation of compositional regression with application to time budget analysis. *Austrian J Stat* 2018; 47(2): 3–19.

20. Dumuid D, Pedišić Ž, Stanford TE, et al. The compositional isotemporal substitution model: a method for estimating changes in a health outcome for reallocation of time between sleep, physical activity and sedentary behaviour. *Stat Methods Med Res* 2017; 962280217737805.

21. Barceló-Vidal C and Martín-Fernández J-A. The mathematics of compositional analysis. *Austrian J Stat* 2016; 45: 57–71.

22. Aitchison J. *The statistical analysis of compositional data*. London: Chapman & Hall, 1986, p. 416.

23. Greenacre MJ. *Compositional data analysis in practice*. London: Chapman & Hall/CRC, 2018, p. 122.

24. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, et al. Isometric logratio transformations for compositional data analysis. *Math Geol* 2003; 35(3): 279–300.

25. Egozcue JJ and Pawlowsky-Glahn V. Groups of parts and their balances in compositional data analysis. *Math Geol* 2005; 37(7): 795–828.

26. Hron K, Filzmoser P and Thompson K. Linear regression with compositional explanatory variables. *J Appl Stat* 2012; 39(5): 1115–1128.

27. Martín-Fernández JA, Pawlowsky-Glahn V, Egozcue JJ, et al. Advances in principal balances for

compositional data. *Math Geosci* 2018; 50(3): 273–98.

28. Chastin SFM, Palarea-Albaladejo J, Dontje M, et al. Combined effects of time spent in physical activity, sedentary behavior and sleep on adiposity and cardiometabolic health markers: a novel compositional data analysis approach. *PLoS One* 2015; 10(10): e0139984.
29. Carson V, Tremblay MS, Chaput J-P et al. Associations between sleep duration, sedentary time, physical activity, and health indicators among Canadian children and youth using compositional analyses 1. *Appl Physiol Nutr Metab* 2016; 41(6 Suppl 3): S294–302.
30. Breslow N. Covariance analysis of censored survival data. *Biometrics* 1974; 30(1): 89-99.
31. Efron B. The efficiency of Cox's likelihood function for censored data. *J Am Stat Assoc* 1977; 72(359): 557–565.
32. Harrell FE and Lee KL. Verifying assumptions of the Cox proportional hazards model. In: *Proceedings of the Eleventh Annual SAS Users' Group International Conference*. North Carolina: SAS Institute Inc., 1986, pp. 823–828.
33. Kleinbaum DG and Klein M. Evaluating the proportional hazards assumption. In: *Survival analysis. statistics for biology and health*. New York: Springer, 2012, pp. 161-200.
34. Grambsch PM and Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994; 81(3): 515-526.
35. Therneau TM and Grambsch PM. Testing proportional hazards. In: *Modeling survival data: extending the Cox model*. New-York: Springer, 2000, pp. 127–152.
36. Rivera-Pinto J, Egozcue JJ, Pawlowsky-Glahn V, et al. Balances: a new perspective for microbiome analysis. *mSystems* 2018; 3(4): e00053-18.
37. R Core Team. R: A language and environment for statistical computing. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2014.
38. Therneau TM and Grambsch PM. *Modeling survival data: extending the Cox model.* New York: Springer, 2000, p. 350.

39. Tang Y, Horikoshi M and Li W. ggfortify: unified interface to visualize statistical results of popular R packages. 2016. *R J* 8(2): 478–489.

40. Hamilton NE and Ferry M. ggtern: Ternary diagrams using ggplot2. *J Stat Soft* 2018; 87(3): 1–17.

41. Troiano RP, Berrigani D, Dodd KW, et al. Physical activity in the United States measured by accelerometer. *Med Sci Sport Exerc* 2008; 40(1): 181–188.

42. Fishman EI, Steeves JA, Zipunnikov V, et al. Association between objectively measured physical activity and mortality in NHANES. *Med Sci Sports Exerc* 2016; 48(7): 1303–1311.

43. Palarea-Albaladejo J, Martín-Fernández JA. A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Comput Geosci* 2008; 34(8): 902–917.

44. Buman MP, Winkler EAH, Kurka JM, et al. Reallocating time to sleep, sedentary behaviors, or active behaviors: associations with cardiovascular disease risk biomarkers, NHANES 2005-2006. *Am J Epidemiol* 2014; 179(3): 323–334.

45. Global recommendations on physical activity for health. Report, World Health Organization, Geneva, 2010.

46. Füzéki E, Engeroff T and Banzer W. Health benefits of light-intensity physical activity: a systematic review of accelerometer data of the National Health and Nutrition Examination Survey (NHANES). *Sports Med* 2017; 47(9): 1769–1793.

47. Martín-Fernández J-A and Palarea-Albaladejo J. Dealing with zeros. In: Pawlowsky-Glahn V and Buccianti A (eds) *Compositional data analysis: theory and applications*. Chichester: John Wiley & Sons, Ltd, 2011. pp. 47–62.

48. Martín-Fernández J-A, Hron K, Templ M, et al. Model-based replacement of rounded zeros in compositional data: classical and robust approaches. Comput Stat Data Anal 2012; 56(9): 2688–2704.

49. Palarea-Albaladejo J, Martín-Fernández JA. zCompositions - R package for multivariate imputation of left-censored data under a compositional approach. *Chemom Intell Lab Syst*

2015; 143: 85–96.

Figure 1

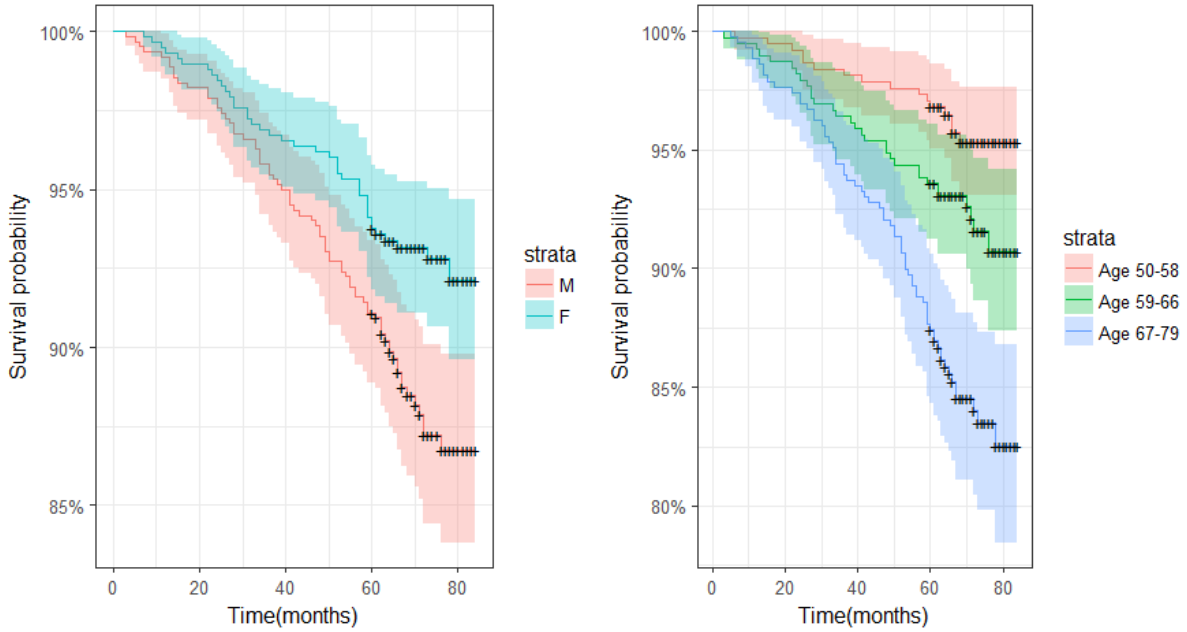

100%
95%
90%
85%
Survival probability
0
20
40
60
80
Time(months)
strata
M
F
100%
95%
90%
85%
80%
Survival probability
0
20
40
60
80
Time(months)
strata
Age 50-58
Age 59-66
Age 67-79

Figure 2

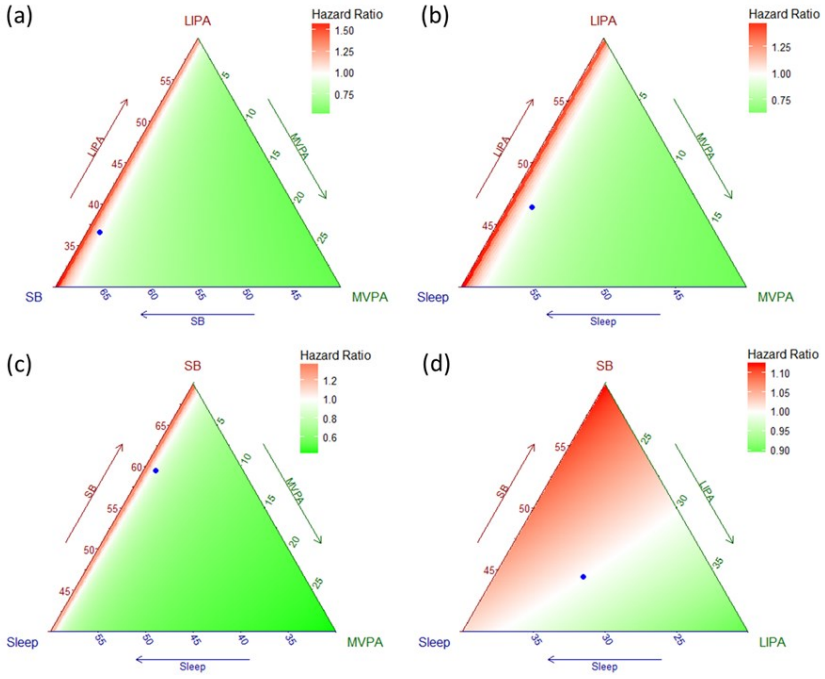
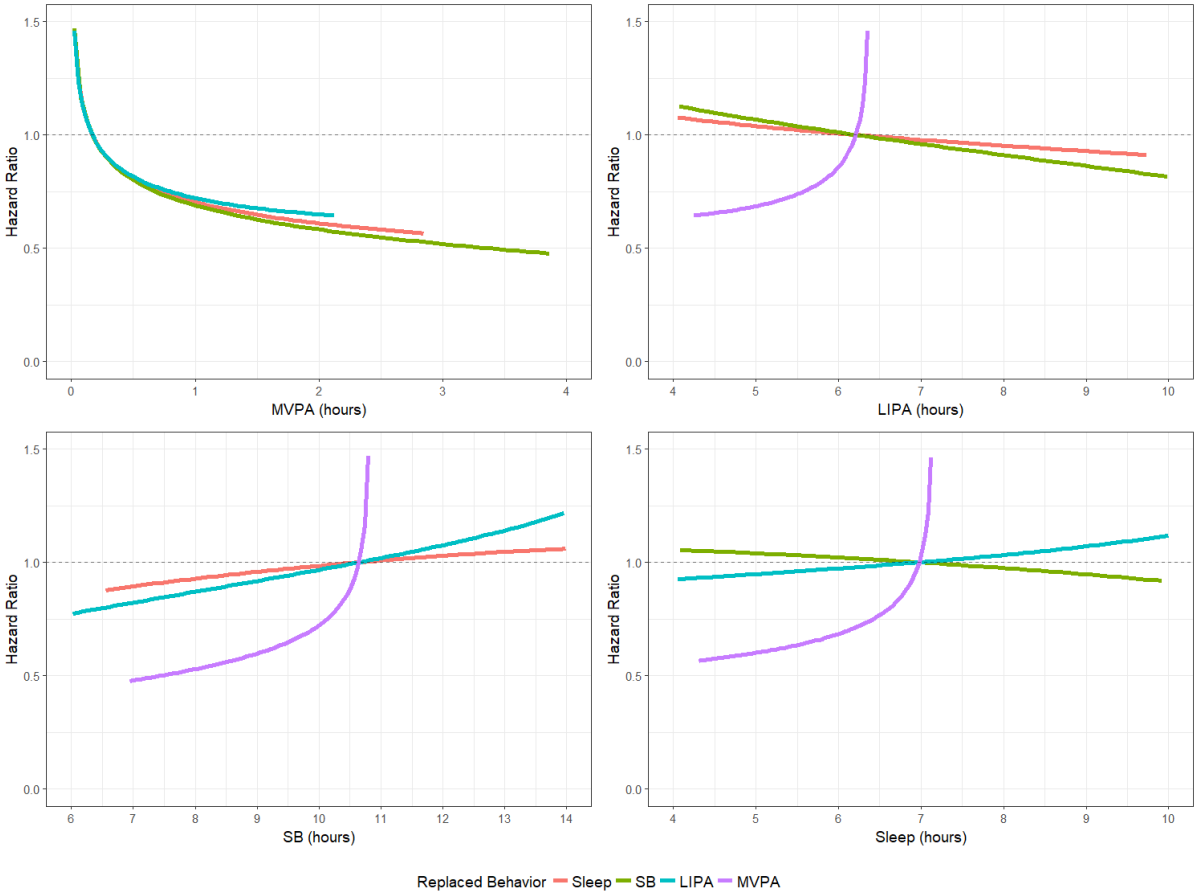

(a)
LIPA
Hazard Ratio
1.50
1.25
1.00
0.75
MVPA
SB
SB
(b)
LIPA
Hazard Ratio
1.25
1.00
0.75
MVPA
Sleep
Sleep
(c)
SB
Hazard Ratio
1.2
1.0
0.8
0.6
MVPA
Sleep
Sleep
(d)
SB
Hazard Ratio
1.10
1.05
1.00
0.95
0.90
LIPA
Sleep
Sleep

Figure 3


Hazard Ratio
MVPA (hours)
LIPA (hours)
SB (hours)
Sleep (hours)
Replaced Behavior
Sleep
SB
LIPA
MVPA

Supplementary Materials: Constructing a basis using Sequential Binary Partitions

1. Basis 1

We set out a process for generating an orthonormal basis based on a sequential binary partition as envisaged by Egozcue at al [ref] using the example of physical activity data. We can envision a sequential binary partition using a simple dendrogram as in Figure S1.

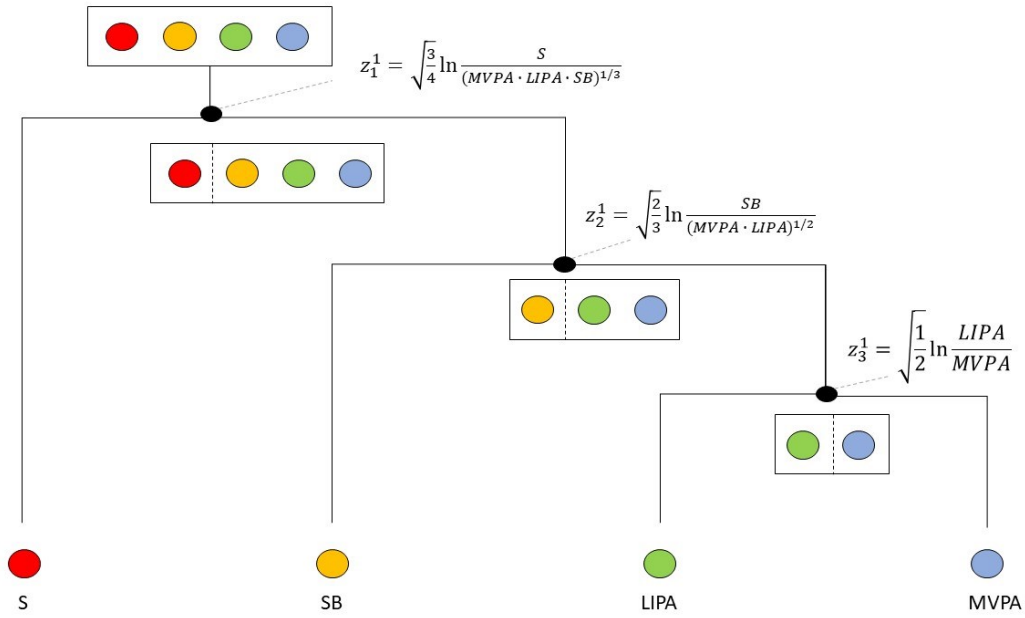


Figure S1: Sequential binary partition underlying Model 1 in the Main Paper

This can be converted to a (D-1) x D matrix **S**. Each row represents a single split. Components not included in the split are set to zero. Components on one side of the split are set to "+1". Components on the other side of the split are set to "-1". It is arbitrary which side is positive so long as both sides are opposite in sign.

S	Sleep	SB	LIPA	MVPA
ilr1	+1	-1	-1	-1
ilr2	0	+1	-1	-1
ilr3	0	0	+1	-1

For each row i we define:

$$a^+ = \frac{+1}{r}\sqrt{\frac{r \cdot s}{r+s}}$$
$$a^- = \frac{-1}{s}\sqrt{\frac{r \cdot s}{r+s}}$$

where r is the number of +1's in row i, and s is the number of -1's in row i.

M	Sleep	SB	LIPA	MVPA	r	s	a^+	a^-
ilr1	+1	-1	-1	-1	1	3	$\sqrt{\frac{3}{4}}$	$\frac{-1}{3}\sqrt{\frac{3}{4}}$
ilr2	0	+1	-1	-1	1	2	$\sqrt{\frac{2}{3}}$	$\frac{-1}{2}\sqrt{\frac{2}{3}}$

ilr3	0	0	+1	-1	1	1	$\sqrt{\frac{1}{2}}$	$-\sqrt{\frac{1}{2}}$

We then replace each "+1" with a^+ and each "-1" with a^- giving the new matrix **M**.

M	Sleep	SB	LIPA	MVPA
ilr1	$\sqrt{\frac{3}{4}}$	$\frac{-1}{3}\sqrt{\frac{3}{4}}$	$\frac{-1}{3}\sqrt{\frac{3}{4}}$	$\frac{-1}{3}\sqrt{\frac{3}{4}}$
ilr2	0	$\sqrt{\frac{2}{3}}$	$\frac{-1}{2}\sqrt{\frac{2}{3}}$	$\frac{-1}{2}\sqrt{\frac{2}{3}}$
ilr3	0	0	$\sqrt{\frac{1}{2}}$	$-\sqrt{\frac{1}{2}}$

We calculate the clr transform as described in the paper.

$$\big(clr(\boldsymbol{x})\big)_i = \ln\left(\frac{x_i}{(\prod_{i=1}^{D} x_i)^{1/D}}\right), \qquad i = 1, \dots, D.$$

Here

$$clr(\boldsymbol{x}) = \left[\ln\left(\frac{S}{g(\boldsymbol{x})}\right), \ln\left(\frac{SB}{g(\boldsymbol{x})}\right), \ln\left(\frac{LIPA}{g(\boldsymbol{x})}\right), \ln\left(\frac{MVPA}{g(\boldsymbol{x})}\right)\right]^T$$

where

$$g(\boldsymbol{x}) = (S \cdot SB \cdot LIPA \cdot MVPA)^{1/4}$$

We then pre-multiply $clr(\boldsymbol{x})$ by **M** to obtain the ilr coordinates.

$$\boldsymbol{z}^1 = \left\{\sqrt{\frac{3}{4}}\ln\frac{S}{(MVPA \cdot LIPA \cdot SB)^{1/3}}, \sqrt{\frac{2}{3}}\ln\frac{SB}{(MVPA \cdot LIPA)^{1/2}}, \sqrt{\frac{1}{2}}\ln\frac{LIPA}{MVPA}\right\}$$

2. <u>Basis 5</u>

Using the same approach we can envision an alternative partition where we initially split the composition between active and non-active behaviours as shown in Figure S2.

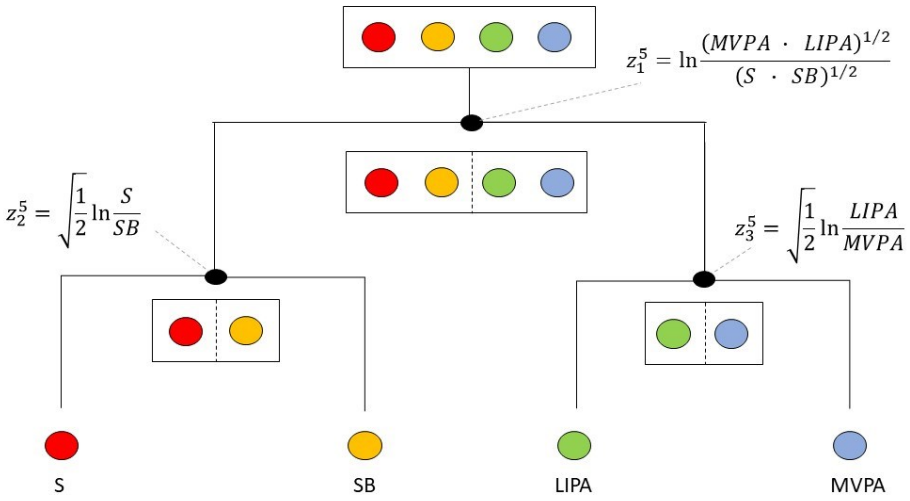


Figure S2: Sequential binary partition underlying Model 5 in the Main Paper

We construct **S** in the same manner.

S	Sleep	SB	LIPA	MVPA
ilr1	-1	-1	+1	+1
ilr2	+1	-1	0	0
ilr3	0	0	+1	-1

We then replace each "+1" with a^+ and each "-1" with a^- giving the new matrix **M**.

M	Sleep	SB	LIPA	MVPA
ilr1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{-1}{2}$	$\frac{-1}{2}$
ilr2	$\sqrt{\frac{1}{2}}$	$-\sqrt{\frac{1}{2}}$	0	0
ilr3	0	0	$\sqrt{\frac{1}{2}}$	$-\sqrt{\frac{1}{2}}$

And again we pre-multiply $clr(\boldsymbol{x})$ by **M** to obtain the ilr coordinates.

$$\boldsymbol{z}^5 = \left\{\ln\frac{(MVPA \cdot LIPA)^{1/2}}{(S \cdot SB)^{1/2}}, \sqrt{\frac{1}{2}}\ln\frac{MVPA}{LIPA}, \sqrt{\frac{1}{2}}\ln\frac{S}{SB}\right\}$$